Nicklaus Choo

Software Engineer · Google

(+1) 669-274-8810 | nchoo@andrew.cmu.edu | www.nicklaus.io | nicklauscyc | Nationality: Singaporean

Work Experience

Google Software Engineer Full-time | Server Load Balancing *SunnyVale, CA, USA* *Aug 29, 2022 - Present*

- Designed and Implemented load balancing for ML/server workloads using GPU/TPU/KV-cache utilization. This achieved:
  - 96% faster Time-to-First-Token P90 latency.
  - 60% faster Normalized-Time-Per-Output-Token P90 latency.
  - 32% faster overall throughput for prefix-heavy workloads.
- Designed and led the implementation of a multi-threaded, concurrent, sharded, distributed load balancing client-server system while working with partner teams in Canada and Poland. This achieved:
  - Cloud customers can balance service traffic along any utilization dimension.
  - Vastly simpler configuration for custom load balancing behavior.
- 10x speedup for network design process with 15x reduction in memory resources for network topology graph data structures.
- Wrote graph traversal algorithms to efficiently traverse Google's network topology and detect single points of failure.
- Improved automation to re-map 25% of all edge network customers to greatly improve customer cost center allocation.

NetApp Software Engineer Intern *SunnyVale, CA, USA* *May 24, 2021 - Aug 20, 2021*

- 3x speedup for OS compile-update-reboot time for ONTAP virtual machines
- Automated ONTAP cloud cluster setup configuration for dynamic load testing
- 2.7x speedup for WAFL scheduler client I/O latency with minimal slowdown (0.8x) to WAFL scheduler replication operations latency
- Further optimized 3.5x speedup for client I/O and 1.2x speedup for replication operations with online random forest server load prediction

ZODAJ Full Stack Software Engineer Intern *Pittsburgh, PA, USA* *May 18, 2020 - Aug 7, 2020*

- Devised SMS and Android COVID-19 contact tracing for Senegal in partnership with Senegalese health authorities
- Designed NoSQL, PostgreSQL databases along with AWS Lambda RESTful API for TCN protocol contact tracing
- Directed and implemented all permissions and roles within ZODAJ for Amazon AWS databases
- Created exposure overlap algorithm for n people in a store that runs in max(O(k) or O(n log n)), where k is overlap count
- Wrote graph algorithms, discrete-time Markov chain SIR epidemic modeling for tracking infections

Noteworthy Projects

32-bit Kernel

- Created x86 kernel from scratch which supports essential syscalls such as fork/exec/wait, pre-emptive multitasking.
- Wrote device drivers for keyboard, timer, and hardware cursor.
- Code available at github.com/nicklauscyc/small-kernel

Education

Carnegie Mellon University *Pittsburgh, PA, USA*

B.S. in Computer Science, Concentration in Algorithms & Complexity with University Honors *Aug 2018 - May 2022*

- Cumulative GPA **3.55**, School of Computer Science Dean's List, High Honors F19, F20, F21.
- Selected courses:

10-701 Machine Learning (PhD)
15-410 Operating Systems
15-411 Compiler Design (C++)
15-440 Distributed Systems (Java)
15-451 Algorithm Design & Analysis
15-459 Quantum Computation
15-356 Cryptography
15-210 Parallel Data Structures & Algos
15-213 Introduction to Computer Systems
15-251 Great Ideas in Theoretical CS
15-259 Probability & Computing
15-260 Statistics & Computing
21-241 Linear Algebra
21-259 Calculus 3D
21-301 Combinatorics
21-373 Abstract Algebra
80-413 Category Theory

Technical Proficiencies

Languages

- C | C++ | Python | Java | Bash | SML | HTML | CSS | JavaScript